Gene-set analysis: unraveling genetic mysteries using the power of molten rock

de Leeuw, C.A.

2019

document version
Publisher's PDF, also known as Version of record

Summary

The goal of this thesis was to investigate the shortcomings of existing gene-set analysis methods and the gene-set analysis approach in general, and to provide a solution for them. Central to this project was the development of a new method for gene-set analysis, MAGMA, which was designed to address the statistical and practical issues identified in other methods as well as to serve as a foundation to extend the gene-set analysis framework in general.

We started in **chapter two** with a gene-set analysis of schizophrenia GWAS data, aiming both to study the glial hypothesis for schizophrenia as well as to test the use of bespoke gene sets tailored to the specific phenotype and hypothesis. Three glia-specific gene lists were curated to do so. These were found to be good representations of their glial type, with each showing unique patterns of overrepresentation for Gene Ontology (GO) biological processes known to be associated with that type.

The 79 functional glia gene sets created using these lists were then tested in a gene-set analysis. The results of this analysis suggested a general involvement of astrocyte genes in schizophrenia, and implicated a number of astrocyte-specific and oligodendrocyte-specific functions as well. Many of the associated gene sets were also found to be particular to one glia type: though GO biological processes of those sets were represented for more than one glial type, they were significant for only one. These results therefore lent support to the glial hypothesis, and also show the value of tailoring the gene sets used in such an analysis to the trait being analyzed.

Chapter three marked the publication of MAGMA. The simulations used to evaluate its performance showed that type 1 error rates were well-controlled, both for the different gene analysis models as well as for self-contained and competitive gene-set analysis. Good performance was maintained when analyzing GWAS summary statistics, despite the need to use smaller and not fully matched reference genotype data to estimate linkage disequilibrium (LD) between variants. When applied to the Crohn's disease GWAS data, MAGMA gene-set analysis was also found to perform well compared to other methods, generally detecting more significant gene sets and performing the analysis considerably faster.

In **chapter four** we provided a comprehensive analysis of the statistical properties of gene-set analysis, and a comparison of specific methods. One key result of this was a demonstration of the fundamental shortcomings of self-contained gene-set analysis. We showed it to be unsuitable for performing biological inference, whereas by contrast the competitive approach did not have this weakness. Power simulations for competitive analysis then uncovered the relations between statistical power and different parameters of the data and genetic architecture of the trait. Of particular note was the finding that power does not indefinitely increase with GWAS sample size, but will eventually level off instead.

The simulations also found flaws in several of the competitive gene-set analysis methods evaluated, relating to failures to sufficiently correct for data-level properties such as gene size or to appropriately account for LD between genes. Of the tested methods, only MAGMA and INRICH were found to have well-controlled error rates under all tested scenarios.

In **chapter five** we introduced a number of extensions to the basic gene-set analysis model and provided a workflow for using them, applying this to blood pressure phenotypes to demonstrate their utility. The joint and conditional analyses uncovered considerable confounding of associations due to general confounders, as well as confounding and overlap of associations among the significant gene properties themselves. We found that the 219 marginal associations from the initial, standard gene-set analysis could be explained by just 28 underlying signals. This demonstrated the value of the extended analysis workflow, in its ability to refine results and remove spurious associations.

Novel results were also found in the interaction analyses, in particular in interactions between gene sets and tissue-specific expression. For these gene sets the association was found to be dependent on the expression in a particular tissue, with the effect of the gene set specific to the subset of genes in

the set that were most strongly expressed. In most cases the marginal associations of these gene sets were weak to non-existent, demonstrating that the interaction analysis was able to detect indirect involvement of such gene sets that would not be found using standard gene-set analysis.

In general, many of the results in this chapter represent novel findings in blood pressure genetics, but they do fit existing knowledge of blood pressure biology quite well. This illustrates that the kinds of model extensions introduced in this chapter can offer additional insight into the biology of traits beyond what basic gene-set analysis can provide, and that such novel methods are needed to obtain a complete picture of the biology behind the genetics.